Day #11 (B)
Excursion 4 Tour IV: Objectivity and Model Checking

4.8 All Models are False (SIST p. 296)

> ... it does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization. ... The construction of idealized representations that capture important stable aspects of such systems is, however, a vital part of general scientific analysis. (Cox 1995, p. 456)

A popular slogan in statistics and elsewhere is “all models are false!” Is this true?

Clearly what is meant involves some assertion or hypothesis about the model

- that it correctly or incorrectly represents some phenomenon in some respect or to some degree.

Such assertions clearly can be true.

To declare, “all models are false” by dint of their being idealizations or approximations, is to stick us with one of those “all flesh is grass” trivializations (4.1).

- So understood, all statistical models are false, but we have learned nothing about how they may be used to infer true claims about problems of interest.

The error statistician’s goal in using approximate statistical models is largely to learn where they break down

Their strict falsity is a given. (So why assign it a probability?)

Probable vs approximate: Assigning a probability to a statistical model is very different from asserting it is approximately correct or adequate for solving a problem.

Two main grounds for the "all models are false" charge (p. 297):

1. The statistical inference refers to an idealized and partial representation of a theory or process.
2. The probability model, to which a statistical inference refers, is at most an idealized and partial representation of the actual data generating source.

- Neither precludes the use of these *false* models to find out true things, or to correctly solve problems.
- On the contrary, it would be impossible to learn about the world if we did not deliberately falsify and simplify.

Adequacy for a Problem. George Box, to whom the "all models are wrong" is often attributed, goes on to add "But some are useful" (1979, p. 2).
I'll go further: all models are false, no useful models are true.

Let's say a statistical model is useful by being *adequate for a problem, meaning*

- it may be used to find true or approximately true solutions to it.
- Statistical hypotheses may be seen as conjectured solutions to a problem.

A statistical model is adequate for a problem of statistical inference (a subset of uses of statistical models):

- if it enables controlling and assessing if purported solutions are well or poorly probed and to what degree.
- Through approximate models, we learn about the "important stable aspects" or systematic patterns when we are in the land of phenomena that exhibit statistical variability.

When I speak of ruling out mistaken interpretations of data, I include mistakes about theoretical and causal claims.

Testing Assumptions is Crucial (p. 298)

> "the ability of the frequentist paradigm to offer a battery of simple significance tests for model checking and possible improvement is an important part of its ability to supply objective tools for learning." (Cox and Mayo 2010, p.285).

The severe tester is a worrywart, which makes her an activist

- deliberately reining in some portion of a problem so that it's sufficiently like one she knows how to check.
- assumptions under test are intended to arise only as i-assumptions.
- They're assumptions for drawing out consequences, for possible falsification.

"In principle, the information in the data is split into two parts, one to assess the unknown parameters of interest and the other for model criticism" (Cox 2006 p. 198).

Number of successes in *n* Bernoulli trials is a *sufficient* statistic,

- has a binomial sampling distribution determined by θ, the probability of success on each trial.
- ***If the model is appropriate*** then any permutation of the r successes in n trials has a known probability.
- Because this conditional distribution (X given s) is totally known, it can be used to assess if the model is violated.
- The key is to look at residuals: the difference between each observed value and what is expected under the model. (We illustrate with the runs test in 4.11.)
- It is also characteristic of error statistical methods to be relatively robust to violation.

This Tour continues our journey into solving the problem of induction (2.7).
(SIST p. 299)

Exhibit (xii). Pest control. Neyman turns from the canonical examples of real random experiments–of coin tossing and roulette wheels–to illustrate how "the abstract theory of probability… may be, and actually is, applied to solve problems of practice importance" such as pest control!

- Given the lack of human control here, he expects the mechanism to be complicated.
- The first attempt to model the variation in larvae hatched from moth eggs, while plausible, is way off.

"[I]f we attempt to treat the distribution of larvae from the point of view of [the Poisson distribution], we would have to assume that each larva is placed on the field independently of the others. This basic assumption was flatly contradicted by the life of larvae as described by Dr. Beall. Larvae develop from eggs laid by moths. It is plausible to assume that, when a moth feels like laying eggs, it does not make any special choice between sections of a field planted with the same crop and reasonably uniform in other respects." (1952, p. 34).

Wrong

- Larvae expert, Dr. Beall, explains why: At each "sitting" a moth lays a batch of eggs.
- "After hatching …the larvae begin to look for food and crawl around" but given their slow movement "if one larva is found, then it is likely that the plot will contain more than one from the same cluster (ibid.)."

- An independence assumption fails.
- The misfit with the Poisson model leads Neyman to arrive at a completely novel distribution: he called it the Type A distribution (a “contagious” distribution.)
- Neyman knows that even the Type A distribution is strictly inadequate, and a far more complex distribution would be required for answering certain questions.
- Yet it suffices to show why the first attempt failed, and it’s adequate to solving his immediate problem in pest control.

Souvenir(U) *Severity in Terms of Problem-solving*.

The aim of inquiry is finding things out. To find things out we need to solve problems that arise due to limited, partial, noisy and error prone information. Statistical models are at best approximations of aspects the data generating process. Reasserting this fact is not informative about the case at hand. These models work because they need only capture rather coarse properties of the phenomena: the error probabilities of the test method are approximately and conservatively related to actual ones. A problem beset by variability is turned into one where the variability is known at least approximately. Far from wanting true (or even "truer") models, we need models whose deliberate falsity enables finding things out.

Statistical methods are useful for testing solutions to problems when this capability/incapability is captured by the relative frequency with which the method avoids misinterpretations.

If you want to avoid speaking of ‘truth” you can put the severity requirement in terms of solving a problem: A claim *H* asserts a proposed solution to an inferential problem is adequate in some respects.

George Box (1983) “An Apology for Ecumenism in Statistics.”

4.9 For Model-checking, They Come Back to Significance Tests

Why can’t all criticism be done using Bayes posterior analysis...? The difficulty with this approach is that by supposing all possible sets of assumptions are known *a priori*, it discredits the possibility of new discovery. But new discovery is, after all, the most important object of the scientific process (Box, G.E.P., 1983 p. 73).

- Box does not view “induction,” as probabilism in the form of probabilistic updating (posterior probabilism), or any other
- Gelman is a Bayesian who follows in this spirit to some extent

Rather, it requires critically testing whether a model M_i is "consonant" with data, and this, he argues, demands frequentist significance testing.

- Our ability "to find patterns in discrepancies $M_i - y_d$ between the data and what might be expected if some tentative model were true is of great importance in the search for explanations of data and of discrepant events" (Box 1983, p. 57).
- But the dangers of apophenia raise their head.p. 301) "This is the object of diagnostic checks and tests of fit which, I will argue, require frequentist theory [of] significance tests for their formal justification". (ibid.)

Once you have inductively arrived at an appropriate model, the move, on his view, "is entirely deductive and will be called estimation." (ibid., p.56).

- The deductive portion, he thinks, can be Bayesian but the inductive portion requires frequentist significance tests, and statistical inference depends on an iteration between the two.

"A model is only capable of being 'proved' in the biblical sense of being put to the test." (Box and Jenkins 1976, p. 286).

- One might imagine A_1, A_2,….,A_k being alternative assumptions and then computing $Pr(A_i|y)$.
- Box denies this is plausible: to assume we start out with all models precludes the "something else we haven't thought of" so vital to science (p. 73).
- Typically Bayesians try to deal with this by computing a Bayesian catchall "everything else."

Savage recommends reserving a low prior for the catchall (1962), but Box worries that this may allow you to assign model M_i a high posterior probability *relative* to the other models considered.

- "In practice this would seem of little comfort" (ibid., pp. 73-4). For suppose of the three models under considerations the posteriors are .001, .001, .998, but unknown to the investigator a fourth model is a thousand times more probable than even the most probable one considered so far?
- So he turns to frequentist tests for model checking.
- Does it violate the likelihood principle (LP)?

> The likelihood principle holds, of course, for the estimation aspect of inference in which the model is temporarily assumed true. However it is inapplicable to the criticism process in which the model is regarded as in doubt....In the criticism phase we are considering whether, given A, the sample y_d is likely to have occurred at all. To do this we *must* consider it in relation to the *other* samples that could have occurred but did not. (Box 1983, pp. 74-75)

- In conducting secondary inferences (about assumptions), Box is saying, the LP must be violated, or simply doesn't apply.

You can run a simple Fisherian significance test–the null asserting the model assumption A holds–and reject it if the observed result is improbably far from what A predicts.

Box gives the example of stopping rules which don't alter the posterior distribution.

He considers 4 Bernoulli trials: <S,S,F,S>.

- The same string could have come about if *n* = 4 was fixed in advance (Binomial trials),
- or if the plan was to sample until the third success is observed, (Negative Binomial trials),
- The string enters the likelihood ratio the same way $\binom{4}{3}\theta^3(1-\theta)$ and $\binom{3}{2}\theta^3(1-\theta)$ respectively: the coefficients cancel in the ratio
- Box contends, this LP violation is altogether reasonable. "In the criticism phase we are considering whether, given A, the sample is likely to have occurred at all" (p. 75).

My question is: How is this secondary inference qualified? Probabilists are supposed to qualify uncertain claims with probability (e.g., with posterior probabilities or comparisons of posteriors).

- Say you have carried out Box's iterative moves between criticism and estimation, arrive at a model deemed adequate, and infer *H*: model M_i is adequate for modeling data $\boldsymbol{x}_0$.
- It's admitted to be a non-Bayesian frequentist animal, but a long-run behavioristic justification wouldn't seem plausible.

Gelman (about Bayesians):

> ,,,not only were they not interested in checking the fit of the models, they considered such checks to be illegitimate…any Bayesian model necessarily represented a subjective prior distribution and as such could never be tested. The idea of testing and p-values were held to be counter to the Bayesian philosophy. (2011, pp. 68-9)

Gelman rejects traditional Bayesian forms.

- "To me, Bayes factors correspond to a discrete view of the world, in which we must choose between models A, B, or C" (Gelman 2011, p. 74) or a weighted average of them (Madigan and Raftery 1994).

- Nor will it be a posterior. “I do not trust Bayesian induction over the space of models because the posterior probability of a continuous parameter model depends crucially on untestable aspects of its prior distribution” (ibid., p. 70).

What is the status of the inference to the adequacy of the model?

- If neither probabilified nor Bayes ratioed, it can at least be well or poorly tested.
- In fact, he says: “This view corresponds closely to the error-statistics idea of Mayo (1996).” (Gelman 2011, p. 70)

4.11 Philosophy of Misspecification (M-S) Testing in the Error Statistical Account

I tell the story of a case Aris Spanos presented to me in 2002.

Nonsense Regression

Suppose that in her attempt to find a way to understand and predict changes in the U.S.A. population, an economist discovers an empirical relationship that appears to provide almost a 'law-like' fit:

$$y_t = 167 + 2x_t + \hat{u}_t,$$

where $\boldsymbol{y}_t$ denotes the population of the USA (in millions), and x_t denotes a secret variable whose identity he would not reveal until the end of the analysis.
The subscript t is time.

There are 33 annual data points for the period 1955-1989 (t = 1 is 1955, t=2, 1956, etc.)
The data can be represented as 33 pairs $\mathbf{z}_0 = \{(x_t, y_t), t = 1,2,\ldots 33\}$. The coefficients 167 and 2 come from the least squares fit, a purely mathematical operation.

This is an example of fitting a *Linear Regression Model* (LRM), which forms the backbone of most statistical models of interest:

$$M_0: \quad y_t = \beta_0 + \beta_1 x_t + u_t, \quad t=1,2,\ldots,n$$

$\beta_0 + \beta_1 x_t$ is viewed as the *systematic* component (and is the expected value of y_t), and
$u_t = y_t - \beta_0 - \beta_1 x_t$ is the error or *non-systematic* component.

The error u_t is a random variable assumed to be Normal, Independent and Identically Distributed (NIID) with mean 0, variance σ^2.

This is called Normal white noise. Figure 4.2 (p. 309) shows what NIID looks like

A Primary Statistical Question*: *How good a predictor is x_t?

The goodness of fit measure of how well this model "explains" the variability of y_t, R^2=.995, an almost perfect fit.

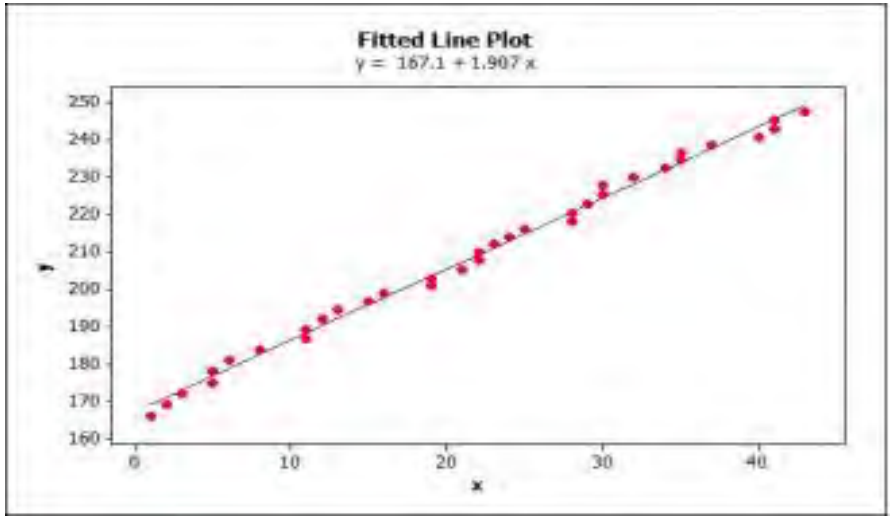


Figure 4.3

The null hypotheses in M-S tests take the form:

H_0: the assumption(s) of statistical model M hold for data ***z***,

as against not-H_0, all of the ways one or more of its assumptions can fail.

To reign in the testing, we consider specific departures with appropriate choices of test statistic d(**y**).

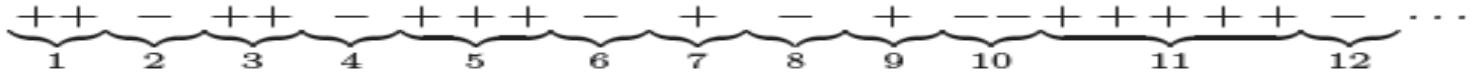
Residuals are the Key

Testing randomness: The non-parametric *runs test* for IID (it falls under "omnibus" tests in Cox's taxonomy, Excursion 3).

Look at the graph of the residuals (Fig 4.4, p. 311), where the "hats" are the fitted values for the coefficients:

$$\{\hat{u}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t, \quad t = 1, 2, \ldots, n\}$$

Instead of the value of each residual, record if the difference between successive observations is positive (+) or negative (-).



- Each sequence of pluses only, or minuses only, is a *run*.
- We can calculate the probability of different numbers of runs just from the hypothesis that the assumption of randomness holds.
- It serves only as an *i*-assumption for the check.

The expected number of runs, under randomness, is $(2n–1)/3$, or in our case of 35 values, 23.

Test statistic: bottom p. 310

The distribution of the test statistic: under IID for $n \geq 20$, can be approximated by N(0, 1).

We're actually testing

H_0: E(R) = $(2n–1)/3$ vs. H_1: E(R) ≠ $(2n–1)/3$.

We reject H_0 iff the observed R differs sufficiently (in either direction) from $E(R)$–23.

Our data yields 18 runs, around 2.4 standard deviation units, giving a P-value of approximately .02.

Arguing from severity, the data indicate non-randomness.

But rejecting the null only indicates a denial of IID: either independence is a problem or identically distributed is a problem: need more specific M-S testing.

The Error in Fixing Error. A widely used parametric test for independence is the Durbin-Watson (DW) test.

Here, all the assumptions of the LRM are retained, except the one under test, independence, which is 'relaxed'.
The original error term is extended to allow for the possibility that the errors u_t are correlated with their own past, *autocorrelated*.

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad t=1,2,\ldots,n,\ldots,$$

This is to propose a new overarching model:

Proposed M_1: $\boldsymbol{y}_t = \beta_0 + \beta_1 x_t + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t,$

(Now ε_t is assumed to be a Normal, white noise process.)

View the D-W test as actually considering the conjunctions:

H_0: {M_1 & $\rho=0$}, vs. H_1: {M_1 & $\rho \neq 0$}.

With the data in our example, the D-W test statistic rejects the null hypothesis (at level .02), which is standardly taken as grounds to adopt H_1.

- This is a mistake: If $\rho = 0$, we are back to the LRM, but $\rho \neq 0$ does not entail the particular violation of independence in H_1.
- we are in one of the "non-exhaustive" pigeonholes ("nested") of Cox's taxonomy.
- Because the assumptions of model M_1 have been retained in H_1, this check had *no chance* to uncover the various other forms of dependence that could have been responsible for $\rho \neq 0$.

Thus any inference to H_1 lacks severity.

The resulting model will *appear* to have corrected for autocorrelation but is in fact statistically inadequate.

What to do instead?

Probabilistic Reduction: Spanos

Spanos shows that any statistical model can be specified in terms of probabilistic assumptions from three broad categories: Distribution, Dependence, and Heterogeneity.

In other words, a model emerges from selecting probabilistic assumptions from a menu of three groups: a choice of distribution; of type of dependence, if any; and a type of heterogeneity

The *LRM* reflects just one of many ways of reducing the set of all possible models that could have given rise to the data $\boldsymbol{z}_0=\{(x_t,y_t),\ t=1,\ldots,n\}$: Normal, Independent, Identically Distributed (NIID).

As a first step, we partition the set of all possible models coarsely:

	Distribution	Dependence	Heterogeneity
LRM	Normal	Independent	Identically Distributed
Alternative (coarse partition)	Non-Normal	Dependent	Non-IID

The *Probabilistic Reduction* (PR) approach to misspecification (M-S) testing weaves together threads from Box-Jenkins, and what some dub the LSE (London School of Economics) tradition.

Rather than give the assumptions by means of the error term, as is traditional, he will specify them in terms of the random variables (x_t,y_t).

This brings out hidden assumptions, notably, assuming the parameters ($\beta_0, \beta_1, \sigma^2$) do not change with t (*t-homogeneity*).

Can indirectly test them from the data.

Clearly, neither data series in Fig 4. 5, 4.6 look like the NIID: the means are increasing with time.

The assumption of linear correlation between *X* and *Y* is that *X* has a mean μ_x, and *Y* has mean μ_y: if these are changing over the different samples, your estimate of correlation makes no sense.

We respecify, by adding terms of form: t, t^2 , …, to the model M_0 to capture the trend

We don't know how far we'll have to go: no inference yet, just building a statistical model whose adequacy for the primary statistical inference will be tested in its own right.

Thus far:

	Distribution	Dependence	Heterogeneity
LRM	Normal	Independent	Identically Distributed
Alternative	?	?	Mean heterogeneity

What about the independence assumption?

- We could check dependence if our data were ID and not obscured by the influence of the trending mean.
- 'subtract out' the trending mean in a generic way to see *what it would be like* without it.

(SIST p. 315)

The detrended data in both figures indicate positive dependence or 'memory' in the form of cycles–Markov dependence.

- So the independence assumption also looks problematic, explaining the autocorrelation detected by the Durbin Watson and runs tests.

- As with trends, dependence comes in different orders, depending on how long the memory is: modeled by adding terms called lags.

- Our assessment so far, just on the basis of the graphical analysis is:

	Distribution	Dependence	Heterogeneity
LRM	Normal	Independent	Identically Distributed
Alternative	?	Markov	Mean heterogeneity

Finally, if we can see what the data $\boldsymbol{z}_0=\{(x_t, y_t), t=1, 2,\ldots, 35\}$ would look like without the heterogeneity ('detrended') and without the dependence ('dememorized'), we could get some ideas about the appropriateness of the Normality assumption.

We do this by subtracting them out “on paper” again.

The scatter-plot of (x_t, y_t), shows the expected elliptical pattern expected for Normality (though I haven’t included a figure).

We can organize our respecified model as an alternative to the LRM.

	Distribution	Dependence	Heterogeneity
LRM	Normal	Independent	Identically Distributed
Alternative	Normal	Markov	Mean heterogeneity

The model derived by re-partitioning the set of all possible models, using the new reduction assumptions of: Normality, Markov and mean-heterogeneity is the Dynamic Linear Regression Model (DLRM).

Back to the Primary Statistical Inference. With a statistically adequate respecified model M_2 we are licensed to make 'primary' statistical inferences about the values of its parameters.

- In particular, does the secret variable to help predict the population of the USA (y_t)?

- No. A test of joint significance of the coefficients of (x_t, x_{t-1}, x_{t-2}), yields a p-value of .823 (using an F test).

- We cannot reject the hypothesis that they are all 0, indicating that *x* contributed nothing towards predicting or explaining *y*.

  The regression between x_t and y_t suggested by models M_0 and M_1 turns out to be spurious or nonsense regression.

Drop the *x* variable from the model and re-estimate the parameters

The secret variable revealed. At this point, Spanos revealed that: x_t was the # of pair of shoes owned by his grandmother over the observation period!

- Some of the best known spurious correlations can be explained by trending means.
- For live exhibits, check out an entire website by Tyler Vigen devoted to exposing them!
- I don't know who collects statistics on the correlation between death by getting tangled in bed sheets and the consumption of cheese, but it's exposed as nonsense by the trending means.

Souvenir (V) *Two more points on M-S tests and an overview of Excursion 4 (p. 317)*